

12:00 - 12:45

Cloudera Academy

“Enseñando tecnología Big Data en el aula”

Dan Johnson, Senior Education Manager EMEA at Cloudera
Ángel García, Product&Service Developer, PUE

cloudera®



cloudera®

Cloudera Academic Partnership

Teaching Hadoop to Tomorrow's Professionals

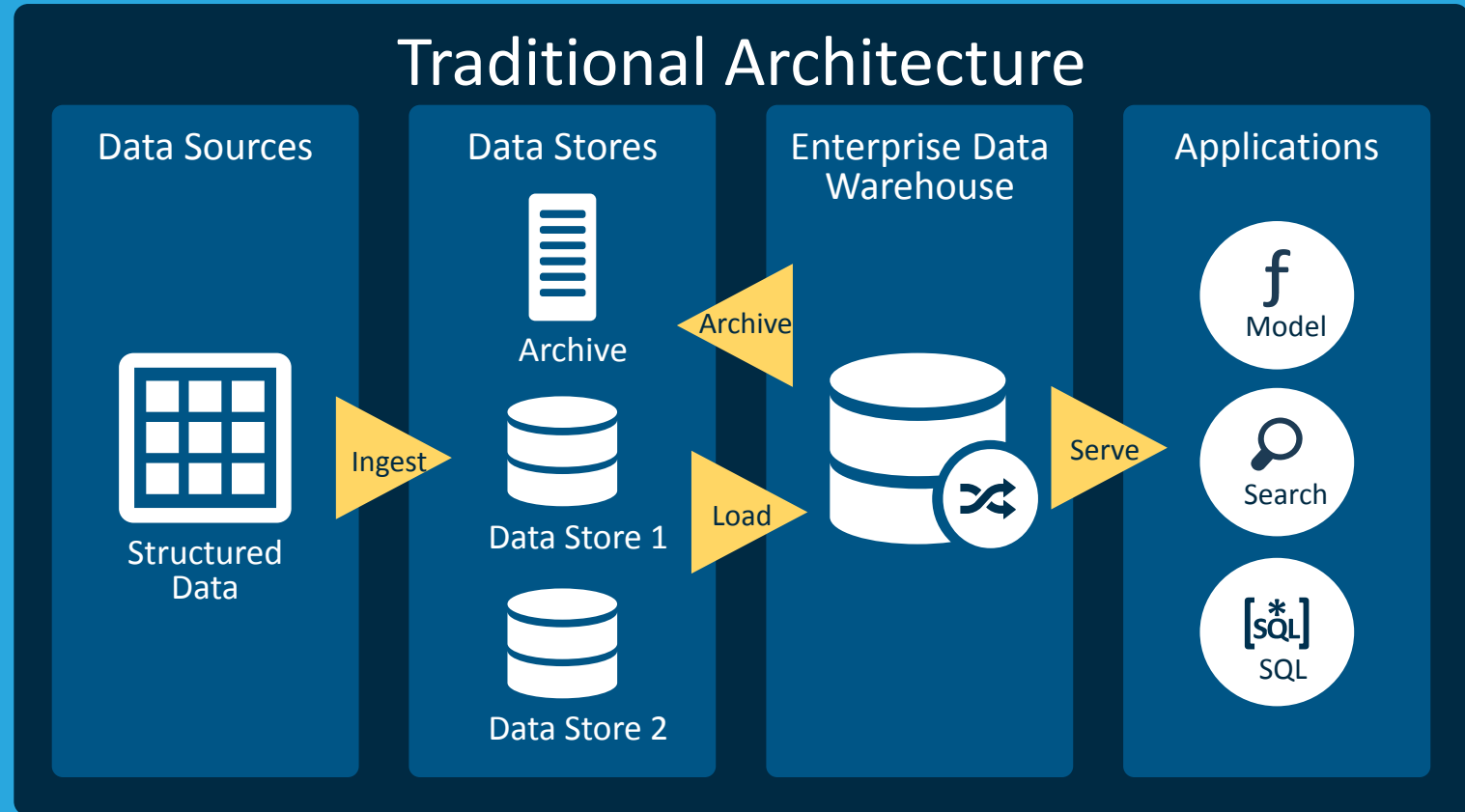


Big data problems & Hadoop solutions

The Old Way

Complex, fragmented, costly

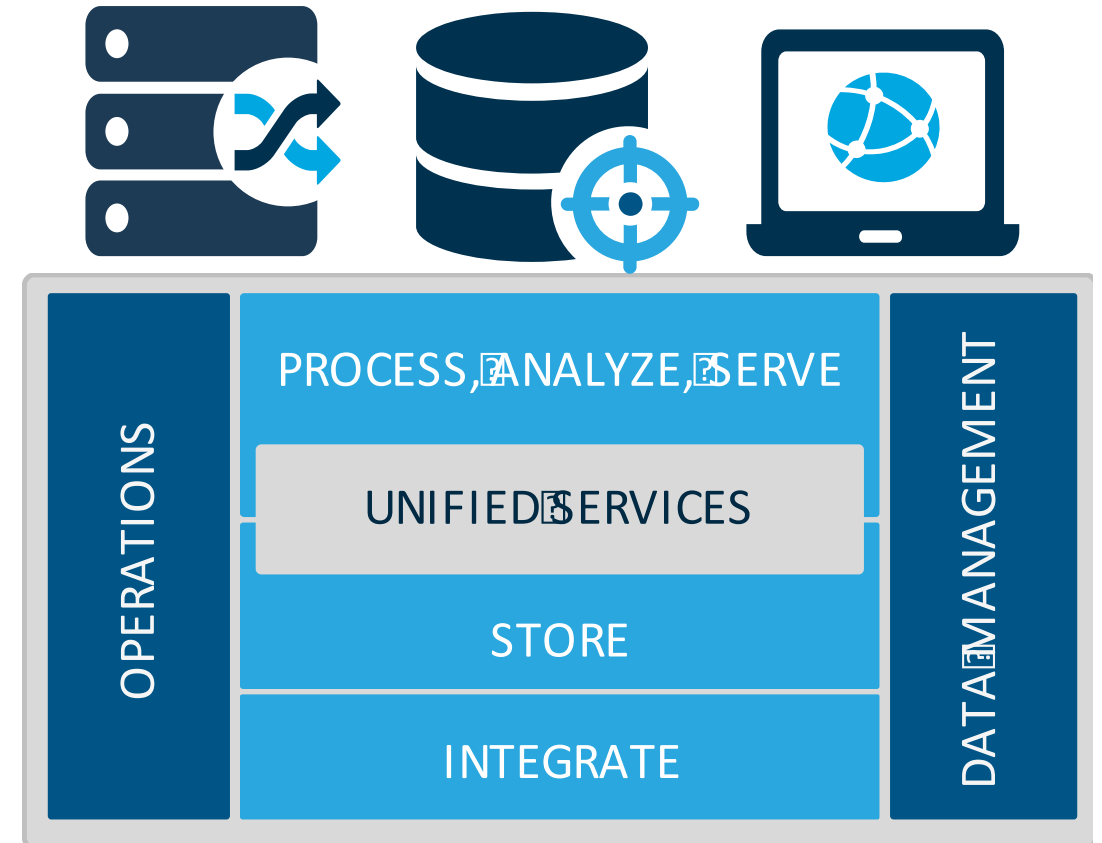
- Data siloed by department or line of business
- Mostly stored in expensive specialized systems
- Analysts only retrieve data into EDW when needed
- Nobody in the organization has a complete view



The Hadoop Way

Simple Unified Efficient

- Data stored on scalable, low-cost platform
- Hadoop handles a variety of workloads
- Perform end-to-end workflows on a single system
- Provides broad data access across departments



Hadoop Powers Big Data Breakthroughs

Delivering True Business Value through Customized Solutions



Financial Services & Banking

Banks ingest, transform, secure, and scrutinize data from multiple silos to detect/prevent fraud and mitigate risk



Healthcare & Biotech

Hospitals index and cross-query patient and pharma data in real time and manage next-gen sequencing workflows



Public Sector & Government

The intelligence community securely manages, audits, and analyzes massive video data sets to fight terrorism



Retail & E-Commerce

Take a 360° view of online behaviors to customize interactions, including real-time recommendations



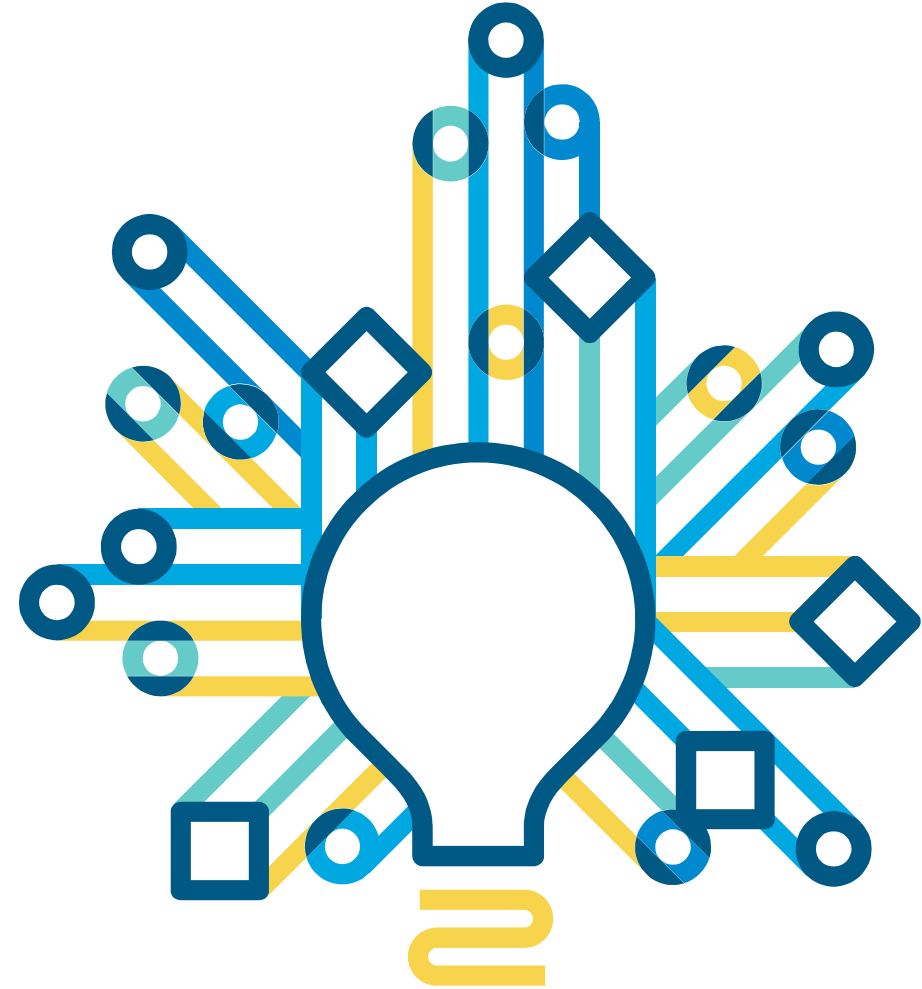
Telecommunications

Mitigate against churn and get a full view of customers to provide targeted marketing and personalization

The data skills gap

Meet the need – big data talent shortage

- The highest paid skills are for Hadoop components or Cloudera, taking up **6 out of the top 10**.
- Hadoop skills are the highest paid in tech and are paid an **average of \$124,927**
- Tech professional are more confident than ever, with **67% saying** that they will be able to find a new position



Global Talent Shortage

The United States, faces a shortage of

140-190k

people with analytical expertise

– McKinsey & Co

80%

of data-intensive UK firms
struggle to find skilled staff

Source: *Model Workers*

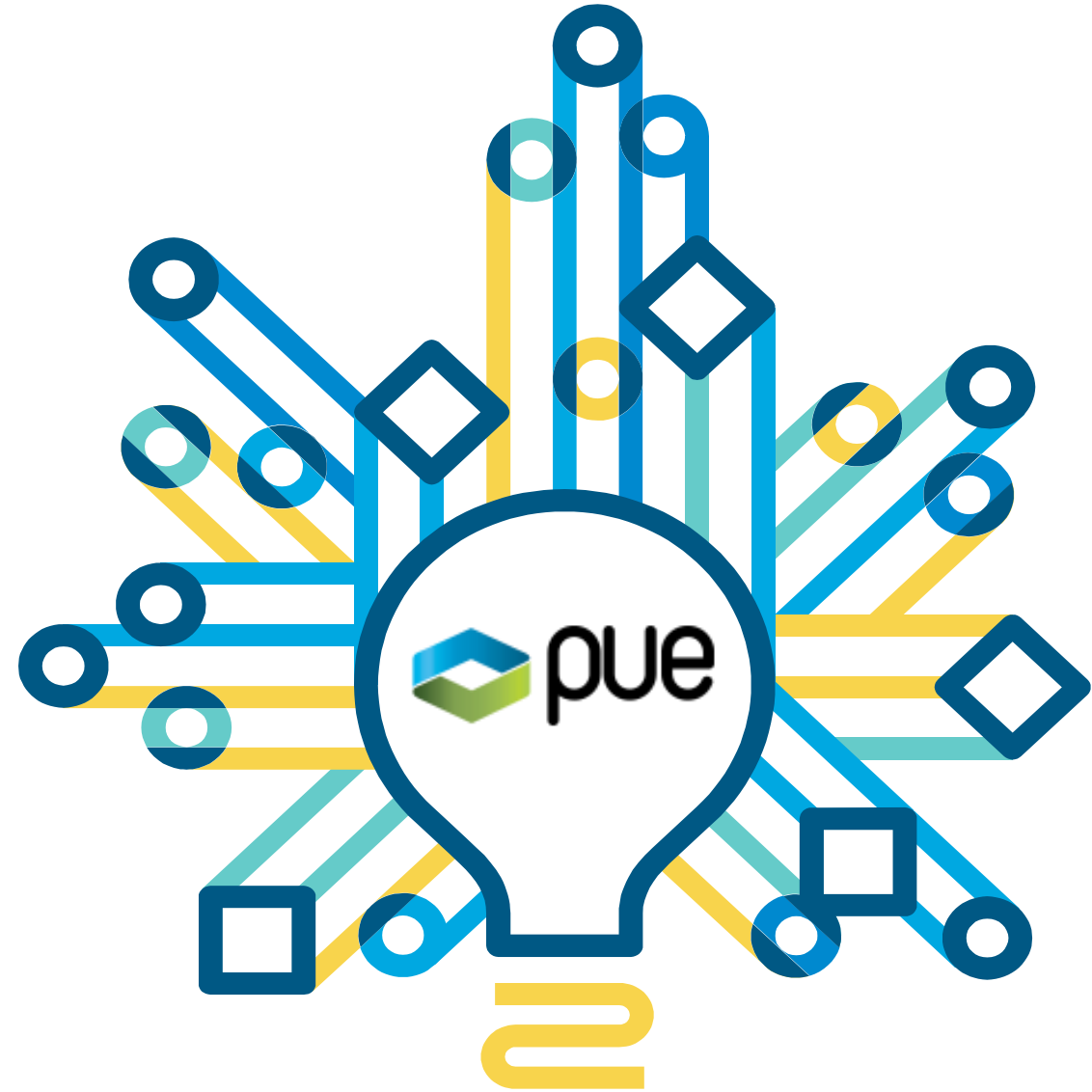
Evolution of the Hadoop Platform

The stack is continually evolving and growing!

											Ibis	Flink
											Mesos	Mesos
											Parquet	Parquet
											Sentry	Sentry
										Spark	Spark	Spark
										Tez	Tez	Tez
										Impala	Impala	Impala
										Kafka	Kafka	Kafka
										Drill	Drill	Drill
						Flume	Flume	Flume	Flume	Flume	Flume	Flume
						Bigtop	Bigtop	Bigtop	Bigtop	Bigtop	Bigtop	Bigtop
						Oozie	Oozie	Oozie	Oozie	Oozie	Oozie	Oozie
						MRUnit	MRUnit	MRUnit	MRUnit	MRUnit	MRUnit	MRUnit
						HCatalog	HCatalog	HCatalog	HCatalog	HCatalog	HCatalog	HCatalog
						Hue	Hue	Hue	Hue	Hue	Hue	Hue
						Sqoop	Sqoop	Sqoop	Sqoop	Sqoop	Sqoop	Sqoop
						Whirr	Whirr	Whirr	Whirr	Whirr	Whirr	Whirr
						Avro	Avro	Avro	Avro	Avro	Avro	Avro
						Hive	Hive	Hive	Hive	Hive	Hive	Hive
						Mahout	Mahout	Mahout	Mahout	Mahout	Mahout	Mahout
						HBase	HBase	HBase	HBase	HBase	HBase	HBase
						ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper
						Solr	Solr	Solr	Solr	Solr	Solr	Solr
						Pig	Pig	Pig	Pig	Pig	Pig	Pig
						YARN	YARN	YARN	YARN	YARN	YARN	YARN
Core Hadoop (HDFS, MapReduce)	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop	Core Hadoop
2006	2007	2008	2009	2010	2011	2012	2013	2014-15				

PUE and Cloudera

- Authorised Cloudera Training Partner for more than 1 year
- Key representative of Cloudera in Spain
- **Leading our Cloudera Academic Partnership (CAP) in Spain for 2016**



Cloudera Academic Program (CAP)

Introducción

La revolución Big Data que genera nuevas necesidades

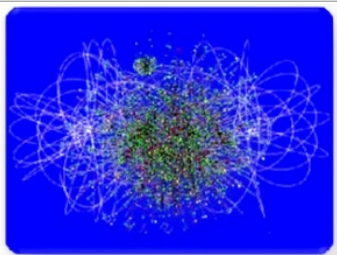
El interés en Apache Hadoop y las soluciones Big Data ha conducido a una demanda creciente de profesionales con habilidades para generar valor a todo el volumen y tipos de datos que se producen.

Un 75% de las empresas invierten o planean hacerlo en Big Data

La tendencia de opinión es que los costes por almacenamiento y la mejora de la experiencia de los clientes son posibles gracias a la inversión en optimización de datos empresariales almacenados.

escrito por: Redacción Computing

18 de marzo 2016



Detrás del **Big Data** se oculta el perfil del cliente, sus hábitos de consumo, sus necesidades y las claves para llegar a ese tan deseado consumidor potencial, pero también se esconde **un 40% de información innecesaria**, que solo una correcta analítica de datos puede detectar, según señala **Wunderman**, compañía de

customer intelligence, CRM, Real Time Marketing y marketing interactivo.

Dificultades para encontrar expertos en análisis de datos

Analytics Trends 2016 de Deloitte indica que la implantación de tecnologías cognitivas permitirá el desarrollo de nuevos puestos de trabajo.

escrito por: Redacción Computing

05 de febrero 2016



La creciente importancia que el **análisis de datos** está adquiriendo en todos los ámbitos de cualquier compañía está contribuyendo de forma significativa a la optimización y transformación de procesos. Sin embargo, la **falta de talento**

capacitado para estas funciones es uno de los grandes retos a los que se enfrentarán en los próximos años. Así lo destaca el **informe Analytics Trends 2016**, elaborado por Deloitte.

más visto

¿Quién utiliza Hadoop?



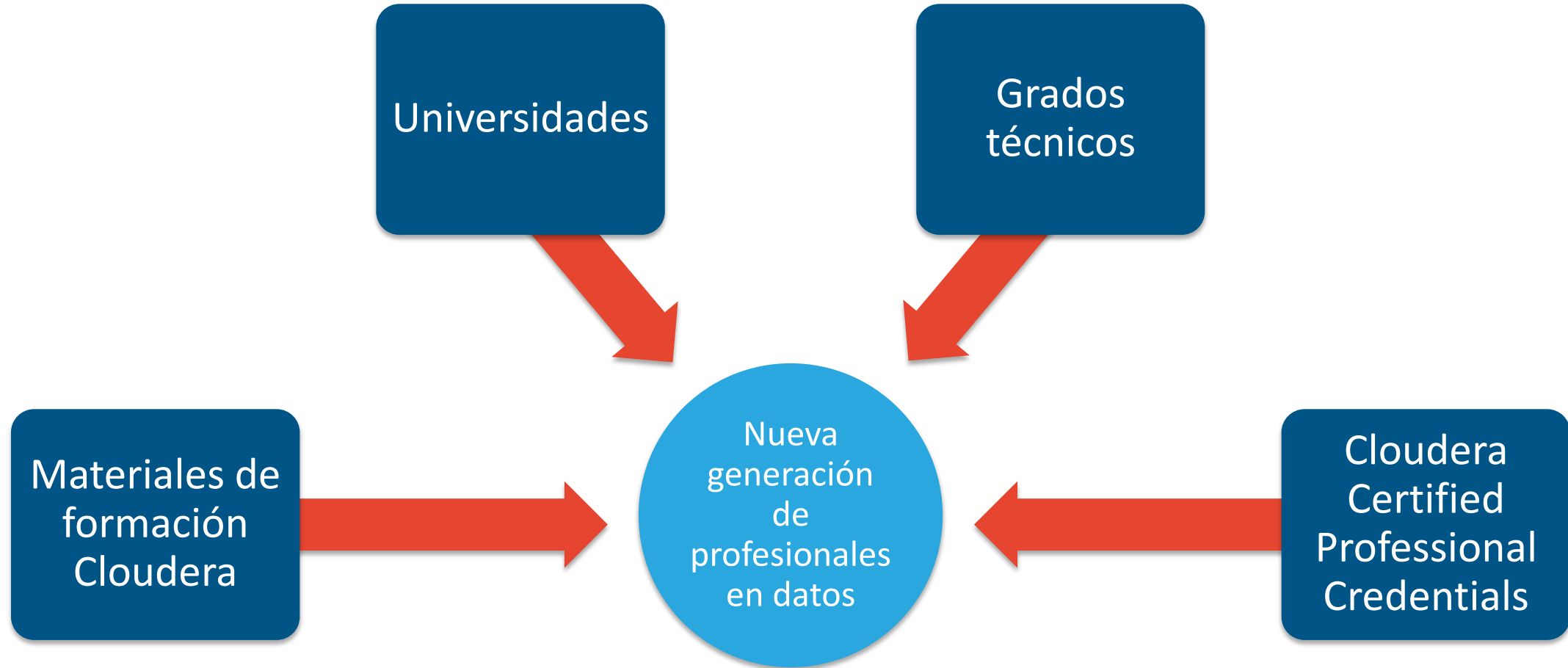
¿Qué es Cloudera Academic Program?

Cloudera Academic Program

- Programa académico de Cloudera que proporciona recursos y materiales para impartir formación oficial en Apache Hadoop
- CAP ha sido creado para ser una relación de colaboración que proporciona beneficios a todas las partes:

Estudiantes	Profesores	Instituciones académicas	Empresas	Cloudera
Acceso a los materiales de formación y certificación más populares de Hadoop, preparándoles para un trabajo en <i>data economy</i>	Guías de instructor, planes de estudio, máquinas virtuales preconfiguradas para ejercicios y descuentos en formaciones y certificaciones	Inclusión de formación Cloudera en el curriculum, proporcionando a los estudiantes habilidades diferenciadoras para encontrar trabajo	Un canal de talento desde instituciones académicas para ocupar necesidades de <i>internships</i> y empleabilidad	Introducir una nueva generación de usuarios de Hadoop, ayudando a asegurar que el mercado tiene el talento para ocupar el crecimiento de Big Data

Cloudera Academic Program



¿Qué ofrece Cloudera Academic Program?

¿Qué ofrece Cloudera Academic Program?

Cloudera Academic Program

cloudera[®]
ACADEMIC PARTNER

Materiales curriculares

Labs y VM

Licencia Cloudera
Manager

Descuentos en cursos y
certificaciones

Descuentos en libros
O'Reilly Media

Materiales de Márketing

¿Qué ofrece Cloudera Academic Program?

Cloudera Academic Program

cloudera[®]
ACADEMIC PARTNER

Materiales curriculares

Labs y VM

Licencia Cloudera
Manager

Descuentos en cursos y
certificaciones

Descuentos en libros
O'Reilly Media

Materiales de Márketing

Cursos oficiales

1. Materiales curriculares para formación en Cloudera

Developing with Spark
and Hadoop

Introduction to
Hadoop and Big Data

Cursos oficiales

1. Materiales curriculares para formación en Cloudera

Developing with Spark
and Hadoop

Introduction to
Hadoop and Big Data

Cursos oficiales

Curso Developing with Spark and Hadoop

El curso cubre Spark y los elementos clave del ecosistema Hadoop utilizado en el desarrollo de aplicaciones finales para procesar eficientemente Big Data.

Una vez finalizado el curso, los alumnos entenderán los conceptos clave de Spark y Hadoop y aprenderán a aplicar herramientas e Spark y Hadoop en el desarrollo de aplicaciones

Cursos oficiales

Curso Developing with Spark and Hadoop

Duración: 36 horas (aprox.)

Prerrequisitos: Conocimiento en Scala o Python, línea de comandos Linux y SQL

Objetivos del curso:

- Encaje del ecosistema Hadoop con el ciclo de vida del procesamiento de datos
- Distribución, almacenamiento y procesado de datos en un cluster Hadoop
- Uso de Sqoop y Flume para ingerir datos
- Procesamiento de datos distribuidos con Spark
- Modelación de datos estructurado como tablas en Impala y Hive
- Elección del mejor formato de almacenamiento para diferentes patrones de uso de datos
- Mejores prácticas para el almacenamiento de datos

Cursos oficiales

1. Materiales curriculares para formación en Cloudera

Developing with Spark
and Hadoop

Introduction to
Hadoop and Big Data

Cursos oficiales

Curso Introduction to Hadoop and Big Data

El curso cubre la arquitectura Hadoop y el ecosistema de herramientas Hadoop. Estas tecnologías son la fundación del movimiento Big Data, que facilitan la gestión de la escalabilidad y el procesamiento de extensas cantidades de datos.



Cursos oficiales

Curso Introduction to Hadoop and Big Data

Duración: 34 horas (aprox.)

Prerrequisitos: Conocimiento en Python, Java o C/C++, redes y sistemas y Linux

Objetivos del curso:

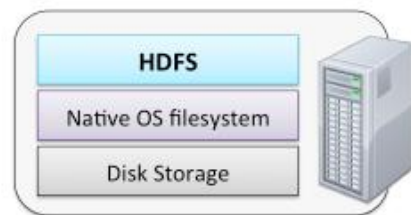
- Necesidad de utilizar Hadoop
- Conceptos del sistema de distribución de ficheros de Hadoop u MapReduce
- Resolución de problemas con Hadoop
- Principales tecnologías de Hadoop y el ecosistema Hadoop
- Desarrollar aplicaciones MapReduce
- Algoritmos frecuentes de MapReduce
- Uso de Pig y Hive para el desarrollo rápido de aplicaciones

Materiales curriculares

Guías del instructor

HDFS Basic Concepts (1)

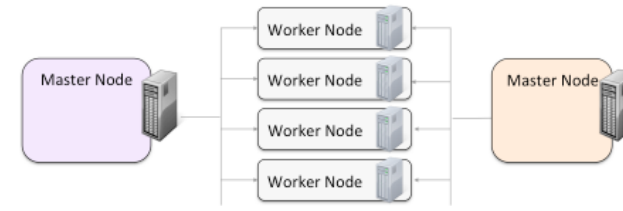
- **HDFS is a filesystem written in Java**
 - Based on Google's GFS
- **Sits on top of a native filesystem**
 - Such as ext3, ext4, or xfs
- **Provides redundant storage for massive amounts of data**
 - Using readily-available, industry-standard computers



cloudera © Copyright 2010-2015 Cloudera. All rights reserved. Not to be reproduced or shared without prior written consent from Cloudera. 3-8

Hadoop Cluster Terminology

- **A cluster is a group of computers working together**
 - Provides data storage, data processing, and resource management
- **A node is an individual computer in the cluster**
 - Master nodes manage distribution of work and data to worker nodes
- **A daemon is a program running on a node**
 - Each Hadoop daemon performs a specific function in the cluster



cloudera © Copyright 2010-2015 Cloudera. All rights reserved. Not to be reproduced or shared without prior written consent from Cloudera. 3-5

A cluster is a group of computers working together – the work is *distributed* across the cluster. As covered previously, distributed processing involves distributing both the tasks and the data the tasks operate on—and therefore a typical cluster has infrastructure for HDFS (to distribute the data) and whatever your chosen cluster management framework is (to distribute the processing).

Materiales curriculares

Slides para alumnos

What Is Apache Flume?

- **Apache Flume is a high-performance system for data collection**
 - Name derives from original use case of near-real time log data ingestion
 - Now widely used for collection of any streaming event data
 - Supports aggregating data from many sources into HDFS
- **Originally developed by Cloudera**
 - Donated to Apache Software Foundation in 2011
 - Became a top-level Apache project in 2012
 - Flume OG gave way to Flume NG (Next Generation)
- **Benefits of Flume**
 - Horizontally-scalable
 - Extensible
 - Reliable



cloudera © Copyright 2010-2015 Cloudera. All rights reserved. Not to be reproduced or shared without prior written consent from Cloudera. 9-5

Chapter Topics

Capturing Data with Apache Flume

Introduction to Flume

- What is Apache Flume?
- **Basic Flume Architecture**
- Flume Sources
- Flume Sinks
- Flume Channels
- Flume Configuration
- Conclusion
- Homework: Collect Web Server Logs with Flume

cloudera © Copyright 2010-2015 Cloudera. All rights reserved. Not to be reproduced or shared without prior written consent from Cloudera. 9-11

Materiales curriculares

Timings para clases y labs

Cloudera Academic Partnership						
Course: CAP - Developing with Spark and Hadoop						
Course Schedule with Timings						
Week	Lecture Components	Component Duration (hours:min)	Class Duration (hours:min)	Comments	Homework Labs	Homework Duration (hours:min)
Lecture 1 - Modules:						
1	Introduction Introduction to Hadoop					
	Course Overview and Logistics	0:30	0:30			
	1. Introduction	0:30	1:00			
	Break	0:05	1:05			
	2. Intro to Hadoop & Ecosystem (lecture part 1)	0:20	1:25			
	Demonstrate CAP Student VM Installation and Use	0:20	1:45	Topics to discuss: 1. How to download the CAP Student VM. 2. How to download and install VMWare. 3. How to run the VM. 4. Navigating within the VM, issuing Linux commands, issuing: <code>sudo jps</code>	Homework: 1. Install the CAP Student VM. 2. Follow instructions in Homework assignment guide for Chapter 2 - Setup and General Notes. 3. Verify that HDFS is working by issuing from the Linux command prompt: <code>hdfs dfs -ls /</code> 4. Read the Google GFS paper.	0:45
	Homework Discussion	0:20	2:05	Describe homework assignment due by next lecture.	Optional homework: Textbook reading assignment	

¿Qué ofrece Cloudera Academic Program?

Cloudera Academic Program

cloudera[®]
ACADEMIC PARTNER

Materiales curriculares

Labs y VM

Licencia Cloudera
Manager

Descuentos en cursos y
certificaciones

Descuentos en libros
O'Reilly Media

Materiales de Márketing

Labs y VM

Homework: Create and Populate Tables in Impala or Hive

Files and Directories Used in this Homework

Exercise directory: `$DEV1/exercises/impala`

MySQL Database: `loudacre`

MySQL Tables: `device`

Data files (HDFS): `/loudacre/webpage`

In these exercises you will define Impala/Hive tables to model and view data in HDFS.

Note: The accounts data will not be used in this exercise but will in a subsequent exercise.

You may perform this and subsequent exercises in either Impala or Hive. Most of the instructions are the same whichever tool you choose; where the instructions differ, the difference is noted. Following whichever set of instructions you prefer; if you have no preference, we suggest Impala because it is faster.

In this Exercise you will explore the performance effect of caching (that is, persisting to memory) an RDD.

1. Make sure the Spark Shell is running. If it isn't, restart it (in local mode with 2 threads) and paste in the job setup code from the previous exercise.
2. This time to start the job you are going to perform a slightly different action than last time: count the number of user accounts with a total hit count greater than five:

```
pyspark> accounthits\  
  .filter(lambda (firstlast, hitcount): hitcount > 5)\  
  .count()
```

```
scala> accounthits.filter(pair => pair._2 > 5).count()
```

3. Cache (persist to memory) the RDD by calling .
4. In your browser, view the Spark Application UI and select the **Storage** tab. At this point, you have marked your RDD to be persisted, but have not yet

¿Qué ofrece Cloudera Academic Program?

Cloudera Academic Program

cloudera[®]
ACADEMIC PARTNER

Materiales curriculares

Labs y VM

Licencia Cloudera
Manager

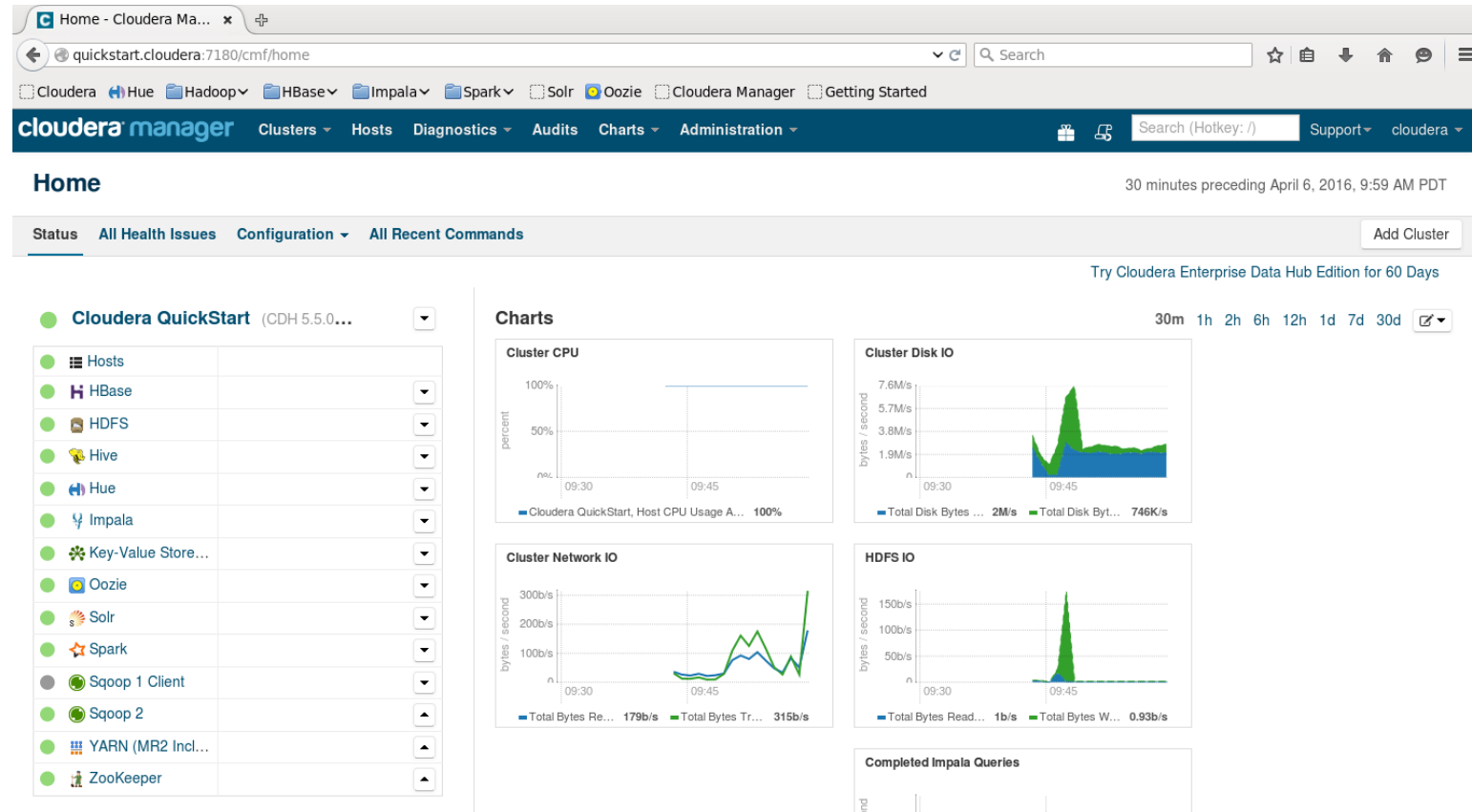
Descuentos en cursos y
certificaciones

Descuentos en libros
O'Reilly Media

Materiales de Márketing

Licencia Cloudera Manager

- Licencia que permite todas las funcionalidades de Cloudera Manager
- No incluye servicios de soporte de Cloudera



¿Qué ofrece Cloudera Academic Program?

Cloudera Academic Program

cloudera[®]
ACADEMIC PARTNER

Materiales curriculares

Labs y VM

Licencia Cloudera
Manager

Descuentos en cursos y
certificaciones

Descuentos en libros
O'Reilly Media

Materiales de Márketing

Precios académicos para miembros CAP

- Descuentos en cursos privados y certificaciones Cloudera

Producto	Descuento aplicable para miembros de CAP
Cursos de calendario impartidos por Cloudera	50%
Cursos privados impartidos por Cloudera	30%
<i>Certification Practice Exams</i>	40%
Exámenes de certification	50%

¿Qué ofrece Cloudera Academic Program?

Cloudera Academic Program

cloudera[®]
ACADEMIC PARTNER

Materiales curriculares

Labs y VM

Licencia Cloudera
Manager

Descuentos en cursos y
certificaciones

Descuentos en libros
O'Reilly Media

Materiales de Márketing

Descuentos libros O'Reilly Media

- Descuentos de hasta un 50% en la compra de libros distribuidos por O'Reilly Media
- 19 títulos disponibles relacionados con temáticas como Spark, Impala, Hive, Pig, etc.

The screenshot shows the O'Reilly Media website with a navigation bar and a main promotional banner. The banner is titled "Welcome Students" and offers a 40% to 50% discount on books in partnership with Cloudera. It lists several books with their original and discounted prices, star ratings, and "Learn more" links. A red "Exclusive" banner is visible in the top right corner of the offer area.

O'REILLY®
Shop Books & Videos

Your Account
Shopping Cart 0 items \$0.00

Home Shop Video Training & Books Radar Safari Books Online Conferences

Browse Subjects | Learning Paths | Video Training | New | Upcoming | Early Release | Bestselling | Ebooks | 1-800-889-8969 / 707-827-7019 / orders@oreilly.com

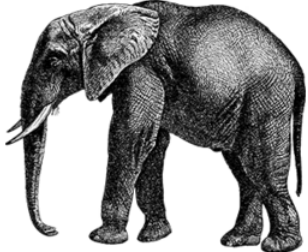
Welcome Students

SAVE 40 to 50%—Your Exclusive Offer in Partnership with Cloudera

Get what you need to thrive in today's science- and technology-driven market. From data collection techniques to visualization and analysis, find it all with books from shop.oreilly.com.

Ebooks from shop.oreilly.com are **DRM-free**. You get **free lifetime access, multiple file formats, and free updates**. Sync with **Dropbox** — your files, anywhere.

Use discount code **CLDR14** - Deal expires March 31, 2016 at 5:00am PT, and cannot be combined with other offers. Offer does not apply to "Print & Ebook" bundle pricing.



Exclusive

Book Title	Original Price	Discounted Price	Rating
HBase: The Definitive Guide	\$42.99	\$21.49	4.0
Learning Spark	\$33.99	\$16.99	4.0
Advanced Analytics with Spark	\$42.50	\$21.25	4.3
Hadoop Application Architectures	\$42.99	\$21.49	4.4
Hadoop Security	\$42.99	\$21.49	-

¿Qué ofrece Cloudera Academic Program?

Cloudera Academic Program

cloudera[®]
ACADEMIC PARTNER

Materiales curriculares

Labs y VM

Licencia Cloudera
Manager

Descuentos en cursos y
certificaciones

Descuentos en libros
O'Reilly Media

Materiales de
Márketing

cloudera



Materiales de Márketing



“Uso del logo oficial como miembro de Cloudera Academic Program”



Materiales de Márketing

“Aparición en el localizador de centros adheridos a Cloudera Academic Program”

CAP Members

North America

Auburn University
California State University, Fullerton
California State University, Los Angeles
Case Western Reserve University
Cinvestav (Mexico)

Texas A&M University
UCSC Silicon Valley
University of Illinois at Springfield
University of Alabama
University of Arizona
University of Bridgeport

EURECOM (France)
Karlsruher Inst. Fur Technologie
KTH Royal Institute of Technology (Sweden)
Nile University (Egypt)
Paris School of Business
Politechnica University of Bucharest

<http://www.cloudera.com/cap>

cloudera



Cloudera & PUE

Cloudera & PUE

PUE ha establecido una relación de colaboración con Cloudera para la gestión e implementación de Cloudera Academic Program en España.

www.pue.es/cloudera-academy

Servicios de PUE

- Formación/videoconferencias (webinars) a los centros sobre la puesta en marcha y utilización de los recursos disponibles.
- Soporte, asesoramiento y resolución de consultas para obtener el mejor rendimiento del proyecto en función de su oferta educativa.
- Recomendaciones y ayuda en la puesta en marcha, implementación y utilización del programa.

Servicios de PUE

- Documentación específica sobre la utilización de los recursos para los centros educativos.
- Envío de información y explicación de las novedades del programa en castellano.
- Gestión directa con Cloudera en el desarrollo y mejora del programa.
- Velar por la correcta utilización del programa

Requisitos y suscripción

Requisitos de participación

- Ser centro de formación reglada homologado para la impartición de Ciclos Formativos de Formación Profesional y Grados Universitarios.
- No aplicar los beneficios y recursos del programa fuera del ámbito de formación curricular reglada del centro

Suscripción

1. Cumplir los requisitos de incorporación al programa
2. Contactar con PUE
3. Abono de la cuota anual de 475,00€ (IVA Exento) en concepto de suscripción y soporte
4. Formalizar la adhesión a Cloudera como Cloudera Academic Partner
5. Recepción de claves de acceso

cloudera



cloudera

[Why Cloudera](#) [Products](#) [Services & Support](#) [Solutions](#) [Get Started](#)

Cloudera Academic Partnership Resources

[Home](#) > [Developers](#) > [Academic Partnership](#) | [Cloudera Academic Partnership Resources](#)

Welcome to the Cloudera Academic Partner Resource Page. Access is restricted to allow only current authorized Cloudera Academic Partners. Here you should find all Cloudera resources requisite to initiating and implementing a successful training program at your institution. If you have any suggestions or comments, please contact actp@cloudera.com

CAP- Developing with Spark and Hadoop

- [Developing with Spark and Hadoop course outline](#) >
- [Course timings – lecture](#) >
- [Course timings – labs](#) >

[Developing with Spark and Hadoop – Instructor Guide](#) +

[Labs for Students – Hints](#) +

[Labs for Students – No Hints](#) +

[Professor VM with Solutions](#) +

[Student VM](#) +

[Student Slides](#) +



cloudera

Thank you

Àngel Garcia

Product & Service Developer at PUE

angel.garcia@pue.es